Social Media Analysis for Multi-faceted Reaction Prediction

Chirag Mandot
109273457
Chirag.mandot@stonybrook.edu

Saipreethy Manickavalli Sayeekrishnan
109231628
smanickavall@cs.stonybrook.edu

ABSTRACT

In the world where prediction of the positivity and negativity of the reviews has become a prominent task, the research is moving towards predicting varied emotions from the text. Our project aims at classification of particular emotion in a news article by trying a range of feature sets. We have used a series of machine learning algorithm ranging from supervised to unsupervised learning and generated a set of outputs. The outputs include the prediction accuracy, the frequency of words, the writing style, and a lot of insights based on the mathematical results. We are taking nine emotion categories into account achieving an accuracy of 79.94%. We have also compared pair of categories using perceptron to analyze what are the factors which marks a difference in prediction. Further, to get a better insight into the text, we performed lexical chain analysis to see prediction distribution i.e. false positives, false negatives, recall, precision etc. and try to understand what might be the reason for confusion among various emotional categories. Overall, we take a news article, predict the emotion with high accuracy and generate a reasoning for that particular output.

1. INTRODUCTION

Emotion detection has become a necessity of research because of its wide range of applications like understanding the nature of a person through his/her writing, a psychologist can treat a patient better if he/she understands the inner emotions in session transcripts, it can help in better human computer interaction devices etc. These problems motivate to analyze the text, find the hidden pattern and using those patterns to predict the outcome.

In this paper, we are analyzing the emotions in news articles. The content in news articles is usually written by well experienced journalist or blogger and they tend to use a set of keywords to induce a particular feeling among the user. For example, they will use specific words for writing a sad articles and another set of words for happy articles. These features can be exploited by applying algorithm and therefore performing sentiment analysis.

There is good amount of work done in this field and multiple approaches have been used. The approaches include supervised, semi-supervised and unsupervised techniques. In this paper we try varied techniques like SVM, Neural Network, KNN, LSA, Clustering, Lexical chain on varied feature sets and compare different techniques, their outputs and corresponding insights obtained.

The emotion detection can be performed in two ways: one is to find the direct words and compare them for every input, another one is to find a pattern, train the classifier and predict based on the learning. The former technique can be done by simple human analysis or simple code. But this approach is not followed because it is too simple and the emotions in the articles are affected by multiple other parameters and not just direct affective words. For example, take the line, "Tom got a new car for his birthday", this line depicts joy but there are not direct words related to it and therefore the simple technique will fail. The solution to this is to have a bunch of articles and use them to train a classifier to predict the output from these hidden patterns.

Another approach of this paper is that running machine learning technique is not enough to obtain high accuracy but a deep understanding of the text and the behavior is necessary. Therefore, this paper has relevant focus both of getting good accuracy and also analyzing the outputs to generate insights from the data.

The rest of the paper is organized as follows. In the next section we present related work done in this domain. Section 3 describes the details of our proposed algorithm. It also describes the results obtained and various techniques tried. Finally, we conclude the paper and discuss some future avenues of research work.

2. RELATED WORK

In this field a range of approaches have been tried and the following are closely related to what we are currently doing in this paper.

One of the first approach [6] used a collection of some 1,336 adjectives and predict based on them. The paper [1] detects sentiments by using two techniques, one is SVM classifier and another is CRF, concluding CRF gives better results as it takes context into account. We have employed SVM by experimenting with features giving us good results. The paper [2] uses combination of varied features like metadata, unigram, bigram to improve accuracy. We have experimented with innovative feature sets and achieved better results comparatively. The paper [3] generates a set of Emotion Generation Rules, designed manually which are used for feature selection. We are employing intuitive and literature survey based techniques for feature selection. The paper [7] focusses on unsupervised learning technique and exploits the co-occurrences of the words. To understand the

data better, we used [8] lexical chains described in this paper. We did few modifications in the technique which are described later in the section. We took inspiration from the previous work and added our own methodology to come up with good technique and therefore, results are explained in next section.

3. METHODOLOGY

The overall method can be seen in the Figure 1. The approach is to generate dataset by crawling, create feature vector from it and apply various algorithms to achieve results and thereby use them to extract insights from them. In few cases use those insights back in the method to increase the efficiency of the technique.

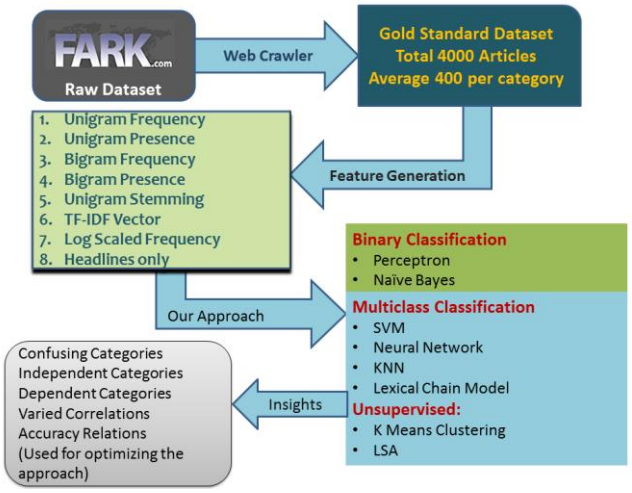

Figure 1: Our Approach

3.1 Dataset

The dataset is crawled from Fark.com using "jsoup" java package which is an html parser and we modified it for the specific purpose. The logic is to take the tag (category of the article) from the metadata on the homepage of the headlines and by following the link of the headline we can extract the content of each news articles by going to their actual news website like NYTimes, DailyNews etc. We quickly parsed through the data to remove extra codes, tags etc. The dirty data might be because the news articles are taken from different websites and the protocol followed by them is not the same for all.

We use 80 percent for training and 20 percent for testing for calculating accuracy. For comparing categories taking two at a time, we use 85 percent for training and 15 percent for testing.

We are using nine news categories for our experimentation and the statistics are as displayed in the table.

Table 1: Dataset description along with some insights into the data

News Category	**# articles**	**Comment about content**
Sad	314	Specific and direct affective words are used
Scary	422	It's an abstract Category and many words which might fall in other categories are used
Strange	353	It consists of few related words and they are majorly used in the articles i.e. frequency of few words dominate the article
Interesting	513	It is also an abstract category therefore, articles do fall in strange or amusing category during classification
Amusing	438	Amusing follows the similar case of interesting category
Dumbass	318	It consist of few words which are very specific and if considered it will give good accuracy
Stupid	331	It contains example and sarcasm majorly
Weird	364	Facts and stories are encountered most of the time
Hero	291	Mostly about achievements or awards

The number of articles vary in each category as there is limited amount of data for some categories. The comment about data is based on the observation we performed before running different algorithms as well as after the output is generated. We proposed various hypotheses and analyzed their truthfulness.

3.2 Baseline

We are using a pretty loose constraint for baseline but it seems to be appropriate considering the problem statement. For every category we take its synonyms and similar words from the link [4], [5]. The occurrence of these words are counted in each text body and based on the maximum count, it is classified. The accuracy obtained is 12.05%.

Table 2 contains the baseline accuracies of different categories.

Table 2: Baseline accuracies as per category

Category	Accuracy (%)
Strange	51.03
Sad	38.77
Amusing	6.8
Stupid	6.36
Scary	4.6
Interesting	1.6

Following are the observations from the baseline experiment:

1. This method will succeed if each article has more direct affective words. But it is not the case; humans are capable of understanding and judging emotions not just based on the usage of words but some intricate details which might not be obvious for the computers to understand.
2. Also, context matters a lot, for ex. the word "new" could mean "interesting" but when combined with words like "new foot prints were discovered" it implies the emotion "scary" as it could mean an alien invasion.

After observing these crucial details, we have employed techniques to overcome these shortcomings and solving as much problem as possible.

Insights are listed in the Table 1.

3.3 Feature Vectors used

Feature vector plays an important role in the results obtained. The paper invests a good amount of time to analyze which features prove to be the best one. In deciding the final feature vectors for classification, many experiments were conducted and the final once which are important enough are as follows:

The feature vectors used are:

1. **Unigram Frequency**
2. **Unigram Presence**
3. **Bigram Frequency**
4. **Bigram Presence**
5. **Unigram Stemming**
6. **TF-IDF Vector**
7. **Log Scaled Frequency**
8. **Headlines only**

We also tried using metadata (author, time, news website etc) for a random sample of data but the results were not improved hence were not extracted for the final dataset.

Insights: Before discussing insights here, it is necessary to understand that the dataset consists of news articles which are written by experienced professionals and very rarely slangs or inappropriate text is observed. Therefore, it's safe to assume a sense of uniformity across the dataset.

The unigram presence gives the best accuracy, which shows that every category has a word pattern which is good enough to classify it correctly. The above argument is justified as unigram frequency decreases the accuracy, the reason could be that some words which are not relevant are getting high weightage, hence diminishing the accuracy. For example, the frequency of the, is, a, an etc. makes the classifier a little biased. After removing these stop words, the accuracy of unigram frequency increased by 2.3% making the conclusion a bit robust.

When the corpus text was stemmed, the accuracy increased by around 3% as the important weights are boosted suppressing the non-important weights

When the frequencies were log scaled, the accuracy was improved a lot as it will normalize the words with high frequency. Therefore, it will normalize the frequency count in the proportion which is perfect for proper classification.

TF-IDF was expected to give good results but it was contradictory after the experiment. As TF-IDF will suppress high frequency words, support rare words and normalize words across dataset, these features might have worked against as follows. Firstly, the high frequency stop words are suppressed which is good but along with it some frequently occurring synonymous words with emotion will also get suppressed. Secondly, the rare words which might not hold much significance will be a playing major role e.g. nouns: words which talk about locations, time, name of the person etc. Giving importance to rare words will be problem because even when we consider articles belonging to same domains they contain articles which talk about different plots and use different words to describe the plot. These words are not common across even the same emotional category, so giving more weightage to them will not help give better results. Lastly, normalization across datasets will suppress strong features which are specific to the dataset.

The headline feature gives a good accuracy because of the following two reasons: first being that headline usually describes the complete article, secondly the word choice is very specific and those words proves to be great feature vector.

3.4 Methods Used

The first step is to calculate accuracy using various methods and observe the results obtained. We perform the evaluation in two phases, once just by considering headlines of the articles as they gives maximum information gain about the article, later taking both headline and body into account for classification and thirdly we will study individual categories to have a deeper insight into the classification and working. We have tried SVM using different feature vectors, as computing accuracy is not the aim of the project, improving it is the goal. So we try to use SVM to obtain the accuracy and then modify it by using the insights observed and thereby improving the accuracy. We have used three classifiers so as to have a general idea why the accuracies differ from classifier to classifier

3.4.1 Headlines

In this approach we took only the headlines into consideration. We employed SVM and did 10 fold cross validation getting surprising accuracy of 60.94 percent. This justifies the hypothesis that headlines are quite self-descriptive and the words used in them are revealing. We think that if numbers of articles are increased then accuracy will increase to a good extent.

The reason we are getting good accuracy is as follows: First of all the words used in headlines are distinct if taken different categories, so classifier can distinguish them easily. Secondly, if data is increased then the word set of a particular category will increase hence making a good classifier. On the other hand, after a certain extent if the dataset is increases it won't give a better accuracy as the words are repetitive and it is not increasing the knowledge base. Graphically it can be shown as:

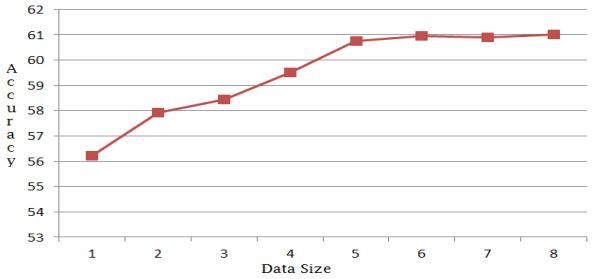

Figure 2: Headline and data size relation

The above graph represents the general idea. We tested it using eight datasets for increasing entries. This opposes the claim which was proposed in initial hypothesis that accuracy will not only become stable but will also decrease but this is not observed. It is possible that if dataset is increased to a great extent then that behavior might be observed. The above test was done in 4000 headline data entries.

3.4.2 Headline and Content

This approach takes both headline as well as content into account. We employ SVM, KNN and Neural Network to classify it. We used the dataset and features described above.

First of all we will used SVM to obtain results. We performed 10 fold cross validation and used trainSVM function in Matlab using linear kernel. Table 3 will describe the results obtained.

Table 3: SVM results

Features	SVM Accuracy %
Unigram Frequency	70
Unigram Presence	79.94
Bigram Frequency	55.1
Bigram Presence	57.9
Unigram Frequency Stemming	73.21
Log Scaled Frequency	73.81
TF-IDF	65
Headlines(Unigram Frequency)	60.94

Insights:

In the above table if we observe the accuracies, it is maximum for unigram presence followed by unigram frequencies. The same pattern doesn't follow for bigram presence as the frequency is less for the combination of the words. Also if the frequency is less, then the pair which is occurring together plays an important role for classification hence giving a better accuracy for bigram frequency. Most of the other insights are shared in the feature vector section above.

3.5 Other Methods

For **Neural Network** we are using Matlab Neural Network Toolkit. The accuracy obtained is not satisfying; therefore a customized neural network needs to be designed. We tried to come up with an algorithm but the accuracy was as low as 22 percent. The customized algorithm was creating memory issues and a lot of time was consumed. We later focused on analyzing the results obtained from SVM.

Table 5: Neural Network Accuracy

Feature	Neural Network Accuracy(%)
Unigram Presence	36.6
Unigram Frequency	31.1
Bigram Presence	30.21
Bigram Frequency	28.75

For **KNN** (K-nearest neighbor), we ran the method for k from 1 to 200 and K=70 was the optimum value. The accuracies obtained from the method are as displayed in the table. We can see that k=70 shows that words from same category are not too close and not too far in the n-dimensional space. If we take lesser K i.e. less number of neighbors then it would give less accuracy because words from other text article are affecting the decision and same is the case if we take K more than 70. It also shows that emotion word clouds are intersecting a lot.

Table 6: KNN Accuracy Table

Feature	KNN(%) K=70
Unigram Presence	67.23
Unigram Frequency	71.22
Bigram Presence	51.11
Bigram Frequency	50.23

Individual Category Analysis:

The individual comparison of the categories and their accuracy will provide a better insight. The accuracies are calculated using SVM and feature vector is unigram frequency. The results are listed for few categories which can help us analyze it critically, considering the fact that excluding other three categories doesn't affect the argument.

Table 7: SVM Accuracy table

Categories	SVM Accuracy (%)
Sad	89.29
Strange	84.64
Amusing	83.33
Interesting	81.13
Stupid	79.22
Scary	74.56
Weird	71.19
Dumbass	69.64
Hero	68.21

The above table provides a nice observation. We see that categories like Sad provides a good accuracy and which is justified as sad emotion can be explained using direct sad words whereas category like scary is difficult to explain using direct words and therefore their explanations might vary a lot from person to person. Also scary is less generalized as compared to sad because sadness can be because of loss in sports, economic market, etc. but scary is pretty extreme and means specific things like death, killing, catastrophe etc. The accuracy of Hero category is low because it consists of heroic activities which are from different fields like a person went to the moon, XYZ swam through the English Channel etc. They use a variety of words hence having a lack of pattern making classification tough.

3.5.1 Calculating confusion amongst categories

In this method we are trying to take two categories at a time and see how similar or different they are. If they are similar then the predicted accuracy should be low when compared to categories that are not so similar or totally different.
We use perceptron to make these comparisons
We compared the following categories:

Similar Categories

Sad and scary
Strange and weird

Dissimilar categories

Interesting and sad
Amusing and sad

Table 8: Comparison of categories using perceptron

Category	Accuracy(%)
Sad and scary	97.14
strange and weird	96.34
Interesting and sad	87.29
Amusing and sad	93.24

After the test, the hypothesis which we predicted is not justified. After careful reading and analysis of around 120 news articles, we could potentially conclude the following reasons responsible:

1. The hypothesis which is made about sad and scary similarity is based on human perception and it is subjective also. But in the gold standard data, the Fark.com people assign different specific domains for both the categories. For example, the scary categories will contain data about terrorist activities, deaths, killings, calamity etc. and the sad category will contain articles like lack of infrastructure, political bribe, sports teams losing, stock market crash etc. If different humans were to annotate these articles then it will be subjective but the Fark.com puts them based on pre-decided domains.
2. Another reason being that the scary articles contains lot of questions like "Is it going to be another 9/11 attack?", "Will there be another hurricane tomorrow?" etc. which is not the case in sad category.
3. The reason for low accuracy between interesting and sad is also counter intuitive. After analysis it seems that many of the articles use the same words but convey opposite meanings for example, "the stock market crashed" (sad), and "the stock market rose high" (interesting), "Arsenal lost by five goals "(sad), and "Arsenal won by five goals"(interesting).
4. Also interesting is mostly expressed using words that could mean some other category like strange, amusing etc. as the article is interesting because it talks about some incident that never happened before. Also it doesn't include words which represent interesting directly.

3.6 Lexical Chain based Approach [10]

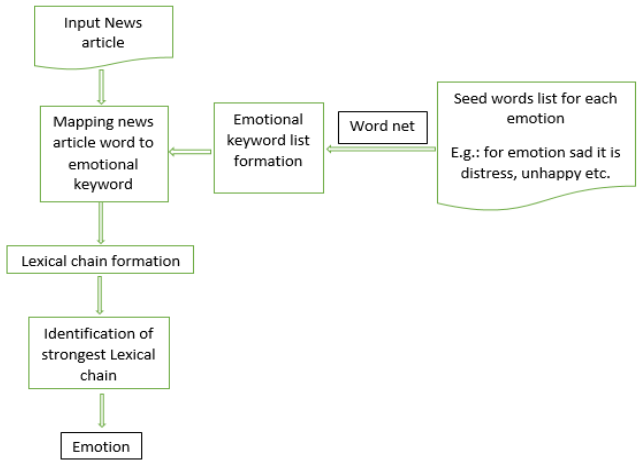

Figure 3: Flowchart showing lexical chain approach

Seed words for each emotional category are found using *Wordnet Synlist*. Emotional keyword list is formed for each emotional category by finding synlist for each of the seed word and adding it to the corresponding emotional keyword list. For e.g. Seed word for the emotional category "sad" includes words like melancholy, distress, etc.

For each word in the news article synlist is found and is mapped to the emotional keyword in the keyword list. Each emotion has a lexical chain which is empty initially. Based on what emotional keyword the element is mapped to, the article word is added to the corresponding lexical chain.

Each lexical chain is given a score based on the length. Strongest lexical chain is the one with the highest score. The article belongs to the emotional category which has the strongest lexical chain.

Figure 3 contains the flow chart which summarizes the overall lexical chain approach.

Table 9: Confusion matrix for lexical chains

	sad	stupid	scary	amusing	strange	interesting
True positive	43	4	18	25	135	152
True negative	1755	1885	1746	1724	843	1156
False positive	243	87	48	54	1111	263
False negative	142	207	371	380	94	612
recall	0.232	0.018	0.046	0.0617	0.58	0.198
precision	0.150	0.043	0.272	0.31	0.108	0.366

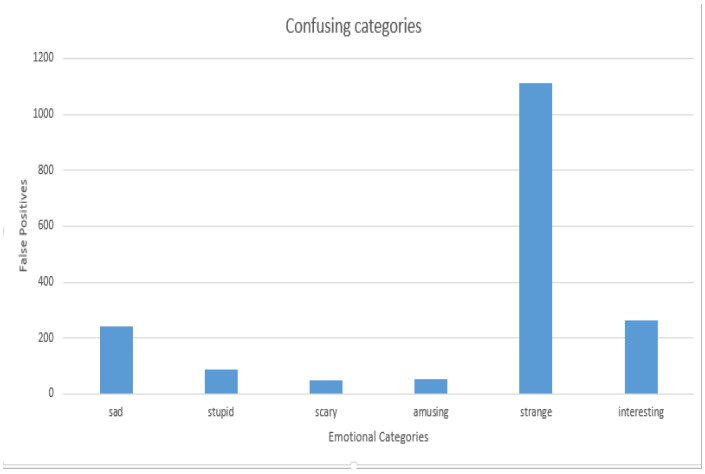

Figure 4: Bar chart showing false positive values

Figure 4 shows that false positive of strange is very high and it also contributes to the confusion.

Insights:

Articles belonging to the category interesting had articles of extreme kinds. Some articles had lot of direct affective words and hence were classified correctly and many were classified wrongly due to lack of direct affective words.

Strange had words overlapping with other categories so, most of the other emotional categories were classified as strange.

Articles belonging to the category stupid did not contain direct affective words nor overlapping words so it lead to low accuracy and did not lead to confusion among other categories also.

Table 9 shows the confusion matrix for lexical chain based approach when 300 articles per category were considered.

Table 10: Confusion matrix after removing confusing category

	stupid	scary	amusing	interesting
True positive	16	75	64	483
True negative	1246	1197	1177	396
False positive	246	127	125	553
False negative	181	290	323	257
recall	0.08	0.205	0.165	0.65
precision	0.06	0.371	0.339	0.466

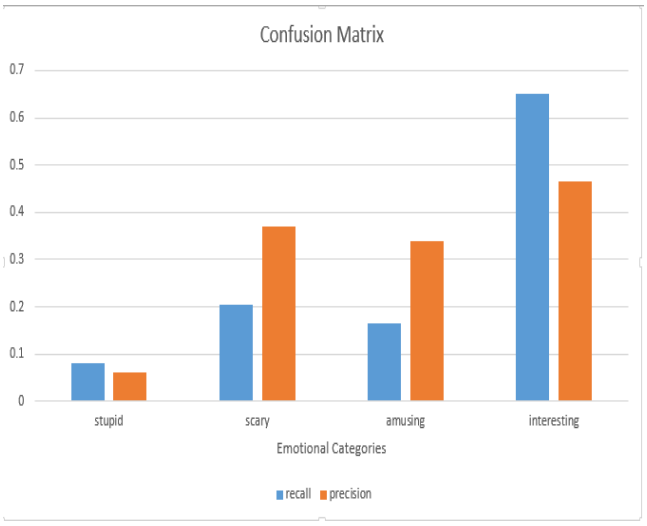

Figure 5: Bar chart showing Recall and precision values

Figure 5 shows the recall and precision values after removing the categories which caused confusion.

When the categories which contained the most overlapping words and caused confusion were removed the accuracy improved by around 17% as shown in Table 10.

To analyze the results better some of the lexical chains were filtered out based on the strength criteria. Each lexical chain is given a score based on their length and homogeneity index. The article has the probability of belonging to the emotional category that satisfies the strength criteria.

Here,

Homogeneity index =1-(# distinct occurrence of each word /length of chain)

Score of lexical chain=Length Homogeneity index.*

*Strength Criterion= Average score of all chains+2*standard deviation of all scores.*

Insights:

During analysis we came across an article which talked about a lady who went to the hotel and threatened all the people there with a knife. She did that because the hot dog they served her was cold. This news actually belonged to the crazy category because she did all this for a hot dog but the plot which describes the incident involved words like knife, threat etc. which lead to the classification of the article into scary category.

3.7 Latent Semantic Analysis:

The main aim of this approach is to find the underlying meaning of a document/news article. It tries to map the words and document together in a "concept space", and this led us to use the technique as it can be helpful for detecting emotions from it. It will also detect the significance of the words occurring together which proves to be important.

We performed the following steps for LSA:

→First of all we computed the count matrix i.e. frequency of occurrence of words.

→We modified count matrix with TF-IDF.

→We performed Singular Value Decomposition using Matlab to have a reduced, accurate, noiseless feature vector.

→Then we compared every emotion category with text in the news article to see the prediction

Results and Analysis from LSA results:

We computed the precision and recall for one of the emotional category. But the results were not outperforming the lexical chain model we explained earlier. We later analyzed the possible reasons for low performance: one of the main reasons is because of the small search query, for example word "sad". This word doesn't give enough details for a proper bootstrapping of the process. Secondly the search query are the general sentiments which doesn't help much because LSA is good technique if the search queries have multiple and unique words

3.8 Unsupervised Approach: Clustering

This approach was used to understand the data in deeper sense. The main reason was to see which categories are completely isolated and which are merging into each other. This analysis can help in projecting the difficulty in classifying a particular category. Our aim in doing unsupervised learning is not to compute accuracy but to make use of unsupervised learning to ease the supervised classification task. It can be explained as:

→When we increase the number of emotion categories to more than 20, then the unsupervised learning plot will help to see which ones are completely isolated and which are not. Based on that graph, the confusing categories will be focused more in the classification task. It will save the amount of computation and the prediction accuracy will boost up.

→Unlike classification task, it will give what is the intensity of separation of data. For example, in classification it will predict a news article as happy or sad, but in clustering using the cosine distance, we can see whether the categories are well separated or just separated.

We used K means clustering which was implemented in Matlab and was ran for 200 iterations. The overall accuracy obtained was 38.41%. The visualization was difficult task the number of dimensions were high and the matrix was very sparse. Therefore the points were closer to origin and selection of right dimension was difficult.

4. CONCLUSION

Emotion detection is one of the most exciting and challenging task. The field requires creative approaches and based on the variety insights and intuitions the prediction accuracy could be improved. It is far more than a normal classification task. In this paper we crawled 4000 articles from Fark.com and cleaned the data. The paper successfully tries to analyze the news articles, individual emotion categories, word usage in them, conclude insights from them and thereby classify them into particular category. We are using eight different types of features vectors. Amongst the classification task SVM performed a good task in classification when feature vector unigram presence was used with accuracy of 79.94%. Other classifier like Neural Network and KNN were ran to conclude the result robustly. The content analysis was done using lexical chain model and the precision and recall of each emotion category were analyzed. A comparison of two categories at a time was performed using perceptron to conclude what marks the difference between them and what factors will lead to boost in prediction results. The paper attempted unsupervised learning for better data visualization and feature selection. An attempt towards dimensionality reduction using SVD was also done. As a future work, the number of categories can be increased and models like CRF/HMM can be tested.

5. ACKNOWLEDGMENTS

Our thanks to Professor Yejin Choi for teaching basic NLP concepts and providing insight into state of the art practices which help us to achieve success in the current project.

6. REFERENCES

[1] Emotion Classification Using Web Blog Corpora, Yang, Changhua ; Lin, K.H. ; Hsin-Hsi Chen

[2] What Emotions Do News Articles Trigger In Their Readers? Kevin Hsin-Yih Lin, Changhua Yang And Hsin-Hsi Chen

[3] Emotion Recognition From Text Using Semantic Labels And Separable Mixture Models, Chung-Hsien Wu, Ze-Jing Chuang, And Yu-Chung Lin

[4] http://www.sba.pdx.edu/faculty/mblake/448/FeelingsList.pdf

[5] http://www.psychpage.com/learning/library/assess/feelings.html

[6] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the

[7] 35th Annual Meeting of the Association for Computational Linguistics, 1997.

[8] Aijun An, Ameeta Agrawal, Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations

[9] Learning to Identify Emotions in Text, Carlo Strapparava, Rada Mihalcea

[10] Emotion Detection using Lexical Chains , M.Naveen Kumar, R.Suresh